D302–D305 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj120

The RCSB PDB information portal for structural genomics

Andrei Kouranov, Lei Xie[1], Joanna de la Cruz, Li Chen, John Westbrook, Philip E. Bourne[1,2] and Helen M. Berman*

Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, [1]San Diego Supercomputer Center, and [2]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Received September 15, 2005; Revised and Accepted October 19, 2005

ABSTRACT

The RCSB Protein Data Bank (PDB) offers online tools, summary reports and target information related to the worldwide structural genomics initiatives from its portal at http://sg.pdb.org. There are currently three components to this site: *Structural Genomics Initiatives* contains information and links on each structural genomics site, including progress reports, target lists, target status, targets in the PDB and level of sequence redundancy; *Targets* provides combined target information, protocols and other data associated with protein structure determination; and *Structures* offers an assessment of the progress of structural genomics based on the functional coverage of the human genome by PDB structures, structural genomics targets and homology models. Functional coverage can be examined according to enzyme classification, gene ontology (biological process, cell component and molecular function) and disease.

INTRODUCTION

The wwPDB (1) maintains the Protein Data Bank (PDB) archives of biological macromolecular structure data, currently comprising over 32 500 structures. Since the year 2000, the worldwide structural genomics initiatives have provided more than 2400 structures, which have also added a large number of new folds. To represent the progress of this collective effort, the RCSB PDB (2) has developed and maintains the Structural Genomics Information Portal at http://sg.pdb.org which consists of three main sections, outlined below.

Structural genomics initiatives

The first section of the information portal provides summary information about each structural genomics center, including target lists, target status, targets in the PDB and sequence redundancy analyses. Summary statistics describing the overall progress of all contributing projects, including sequence similarity and number of structures determined, are regularly tabulated. As an example, an analysis of the sequence similarity of structures solved by structural genomics projects relative to structures in the PDB archive is shown in Figure 1.

Targets

The Targets section offers databases that track target registration data. Currently, 20 structural genomics centers contribute data to the TargetDB (3) resource (http://targetdb.pdb.org). These data include contributing project and target identifier; protein name, source organism and sequence; current production status (e.g. cloned, expressed and crystallized); related database references; and links to related project information. TargetDB assembles data from all contributing centers and makes these data available in a single validated XML data file which is updated weekly.

Targets can also be selected by searching TargetDB by target identifier, similar sequence, program or project, current production status, protein name or source organism. Search results can be captured in FASTA, TargetDB XML or HTML formats. The HTML report presents all of the contributed details about each target including links to related project information and archival databases [e.g. sequence, PDB and BMRB (4)], and links out to protein domain databases. An additional online form constructs cumulative reports summarizing the status of a particular program or project.

Created as an extension to TargetDB, the Protein Expression Cloning and Purification Database, PepcDB (http://pepcdb.pdb.org), was established to collect more detailed status information and the experimental details of each step

*To whom correspondence should be addressed. Tel: +1 732 445 4667; Fax: +1 732 445 4320; Email: berman@rcsb.rutgers.edu

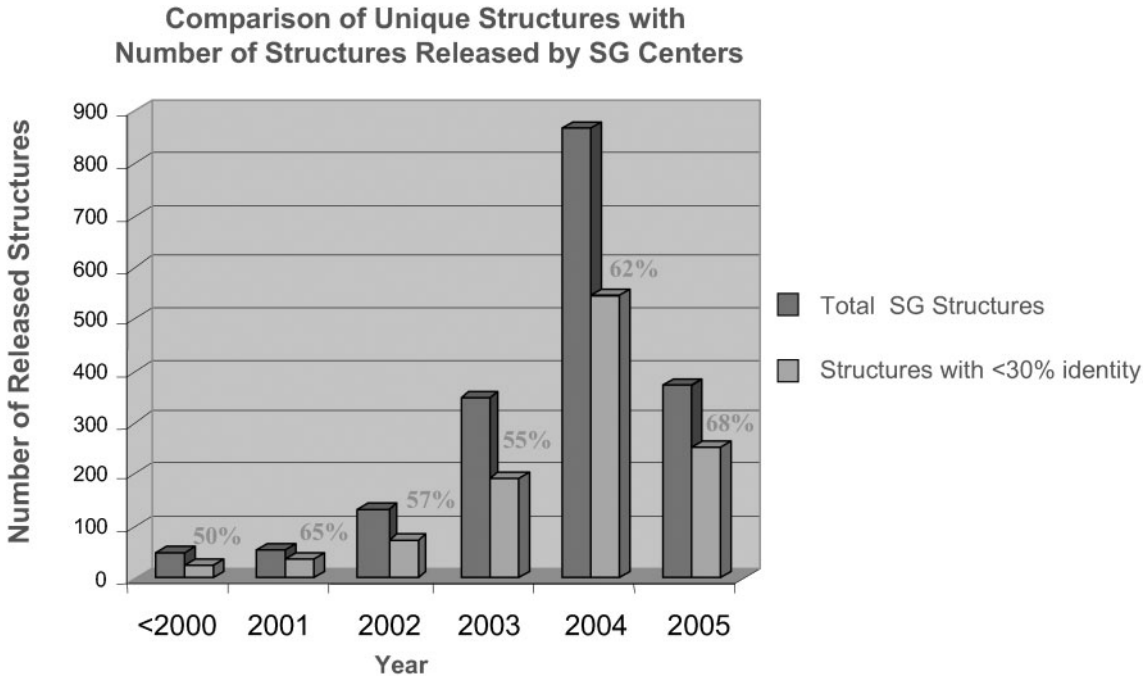

Figure 1. August 2005 report from the structural genomics information portal showing structural genomics structures with sequence similarity <30% relative to solved structures in the PDB by year. Sequence comparisons are performed using the blastclust application (7).

in the protein production pipeline. PepcDB captures a complete history of the experimental steps in each production trial, in addition to describing the current target production status. The status history in PepcDB also records the time interval required to complete each experimental step, with an explanation if work on a particular target or experiment was terminated. Standard protocol descriptions are collected in text form for each step of protein production. Multiple experimental trials can be described for each target. Each trial may reference a set of standard protocols and optionally include the special details of an experimental step and the experimentally observed sequence.

A validation server has been provided for PepcDB contributors (http://pepcdb.pdb.org/validation.html). Data files validated through this form are automatically loaded into the PepcDB database. PepcDB currently includes protocol information from the NIH Protein Structure Initiative (PSI) centers. TargetDB status data from all other structural genomics centers are merged into PepcDB. As a result, PepcDB always provides the most complete view of target status and experimental information for structural genomics projects.

The search features of PepcDB build upon those of TargetDB by offering additional tools to mine experimental protocols. Protocol searches are integrated with queries for target sequence and other target attributes. The resulting report includes the essential target description provided by TargetDB plus additional links to a chronological status history and links to related experimental protocols.

Structures

The Structures section of the RCSB PDB Structural Genomics Information Portal (http://function.rcsb.org:8080/pdb/function_distribution/index.html) provides information about the functional distribution of solved structures, structures being determined by structural genomics and homology models determined from solved structures (5). Function is measured relative to Ensembl-assigned functions from the human genome (6) and disease relative to OMIM assignments for human diseases (7). This section answers the question 'With respect to the function of proteins identified in humans and human disease, what does the present complement of structures in the PDB, the structural genomics targets (if all were solved) and homology models that can be built from the current set of templates add to our understanding of living systems?' The answer to this question changes over time, and the functional distribution resource provides a current answer since the constituent components needed to address the question—PDB structures, structural genomics targets, homology models from SUPERFAMILY (8), functional assignments from Ensembl and disease classifications from OMIM—are all updated as they change, ranging from weekly for PDB structures and targets to approximately annually for SUPERFAMILY. The answer to the question also depends on the definition of a homology model. Here the structural templates used in homology modeling were a set of hidden Markov models taken from SUPERFAMILY 1.65. The sequences were aligned to the structural template with HMMER (9). Only those assigned domains with sequence identity >30% in the alignment were considered as homology models.

Through the functional distribution site, this question can be addressed by examining molecular function, biological process and cellular component [as assigned by the Gene Ontology, GO (10)], enzymes via their EC numbers (www.chem.qmw.ac.uk/iubmb/enzyme), and diseases assigned through OMIM (7). Several steps are used to define the search parameters; here molecular function is used as an example.

In Step 1 (Molecular Function) the breadth of the search is defined, which in turn defines the details presented in the results. So, for example, the top level of the GO hierarchy for molecular function is displayed and used by default. All structures could be selected, or a subgroup could be selected (e.g. all structures with the molecular function 'vitamin binding') by browsing through the hierarchical tree. Similarly in Step 2 (Structure Type), all structures are chosen by default, but it is possible to drill down and explore just groups of structures based on the SCOP classification of class (all alpha, all beta, etc.) (11). Step 3 selects the genome. At present only the human genome is available, but other model organisms will be added. Step 4 selects the sequence identity to use, with 40% identity the default. Sequence identity defines how the human genome sequences are clustered and a single function assigned for that cluster—at lower sequence identity there are fewer clusters, i.e. the results are effectively at lower resolution. Step 5 specifies the domain combinations needed for a match. Since PDB structures frequently represent a single domain in a larger complex, statistics can be produced requiring overlap for one or more domains up to the whole structure accounting for domain rearrangements [see (5) for a full description].

Based on these input parameters, one of three distributions can be generated: a comparison of the distribution of PDB structures, structure genomics targets or homology models against the human genome; a 'most wanted list'

Table 1. Genome coverage

	Function coverage	Cluster coverage
Genome sequences	1.000	1.000
PDB structures	0.372	0.094
SG targets	0.324	0.156
Homology models	0.563	0.283
PDB structures + SG tragets	0.515	0.239
PDB structures + homology models	0.595	0.303
SG targets + homology models	0.663	0.411
PDB structures + SG targets + homology models	0.687	0.428

Data are based upon 10 801 functionally described human genome sequences from Ensembl, 942 PDB structures from human, 1680 structural genomics targets identified in human and 2823 homology models from SUPERFAMILY mapped on to the human genome. Cluster Coverage is the ratio of number of protein clusters that are structurally covered versus all clusters in the genome for a functional class with a specified sequence identity (40% in this case). Functional class and sequence identity are input parameters.

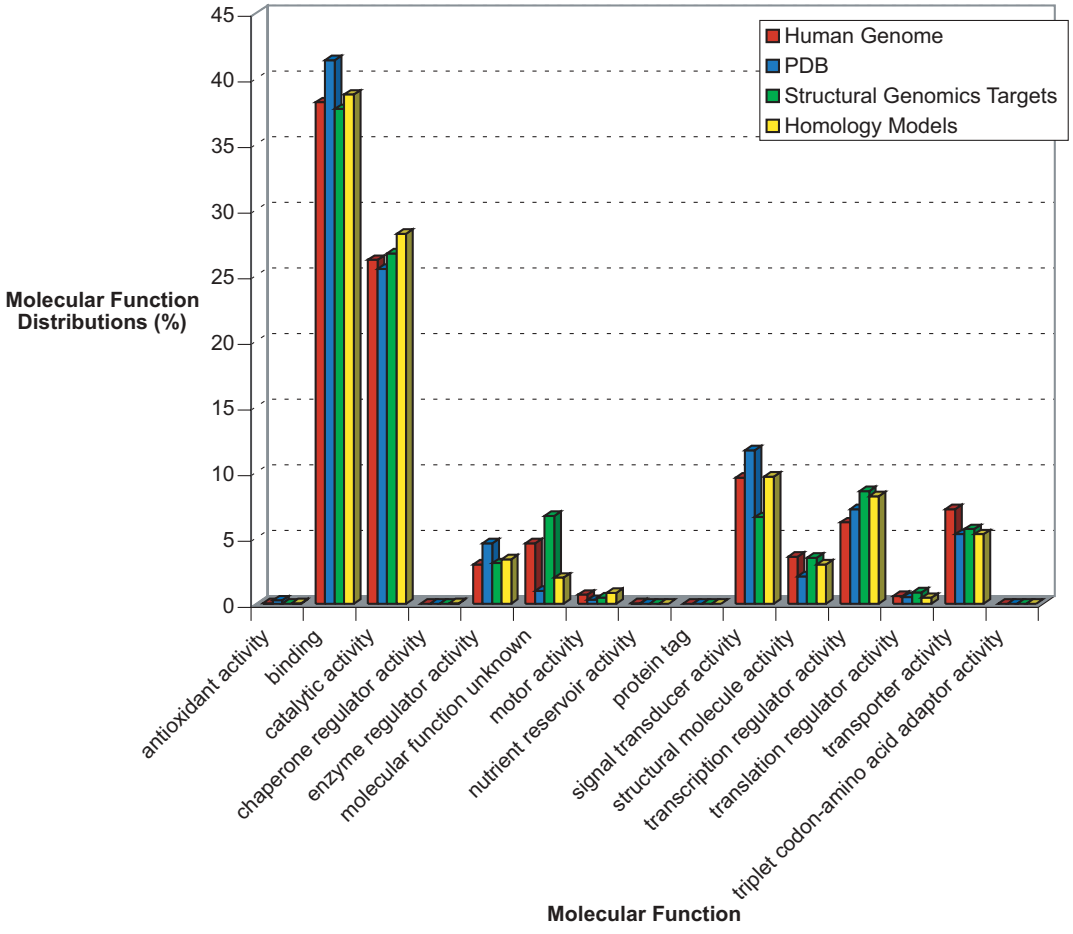

Figure 2. Normalized functional coverage of the human genome by sequence (from Ensembl; red), by structures from the PDB (blue), by structural genomics targets (green) and homology models from SUPERFAMILY (yellow). When viewing the figure from the online structural genomics portal, clicking on the appropriate bar of the histogram will produce a list of sequences or structures that define the distribution.

of structures—those not in the PDB and which (by default) are not identified through homology modeling or in the structural genomics targets yet have significant presence in the human genome; and simple charts showing the distribution of the genome sequences, PDB structures, structural genomics targets or homology models. Most distributions are accompanied by two tables illustrating, first, the functional coverage by each data type (Table 1), and second, the correlation between input data types (data not shown). The actual overlap between these groups will be added as part of an on-going development. For example in Table 1, PDB structures cover 37.2% of the identified molecular functions in the human genome; if solved, structural genomics targets cover 32.4% of functions; and 56.3% of the molecular functions can be modeled from existing structures. Figure 2 illustrates the resulting normalized distributions for the top level of the GO molecular function hierarchy. At this level most distributions are not skewed with the exception of molecular function unknown—PDB structures are underrepresented and structural genomics targets are overrepresented. Not surprising, since until structural genomics began structural biology was dominated by determining structures of known function. In the era of structural genomics, that trend has reversed. Drilling down to more detailed descriptions of molecular function (data not shown) reveals a more uneven distribution and suggests changes in structure determination strategies.

An important feature of this resource is the 'most wanted list' of structures based on the following criteria: (i) functional categories where proteins are underrepresented by structures; (ii) from (i), proteins which can not be modeled, i.e. proteins from the human genome without SUPERFAMILY assignments; (iii) if the protein can be associated with a human disease; and (iv) proteins identified as likely to be intractable, i.e. with a transmembrane segment filtered out.

CONCLUSION

In this report, we present the resources currently made available through the RCSB PDB in support of the structural genomics effort. It is expected that further functionality will be added as the second phase of the PSI and other worldwide efforts move forward.

ACKNOWLEDGEMENTS

The RCSB PDB is operated by Rutgers, The State University of New Jersey and the San Diego Supercomputer Center at the University of California, San Diego. This work is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the National Institute of Neurological Disorders and Stroke (NINDS). The RCSB PDB is a member of the wwPDB. The Structures section of the portal was funded in part by NIH grant GM63208. Funding to pay the Open Access publication charges for this article was provided by the NSF.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.*, **10**, 980.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
3. Chen,L., Oughtred,R., Berman,H.M. and Westbrook,J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, **20**, 2860–2862.
4. Doreleijers,J.F., Mading,S., Maziuk,D., Sojourner,K., Yin,L., Zhu,J., Markley,J.L. and Ulrich,E.L. (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR*, **26**, 139–146.
5. Xie,L. and Bourne,P.E. (2005) Functional coverage of the human genome by existing structures, structural genomics targets, and homology models. *PLoS Comput. Biol.*, **1**, e31.
6. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
7. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
8. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235–D239.
9. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
10. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
11. Lo Conte,L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.